Substitution of hazardous chemical substances using Deep Learning and t-SNE

Patrick Lübbecke, Nijat Mehdiyev, Peter Fettke

German Research Center for Artificial Intelligence, Institute for Information Systems, Saarbrücken, Germany
{patrick.luebbecke, nijat.mehdiyev, peter.fettke}@dfki.de

Abstract. Manufacturing companies in the European Union are obliged to regularly analyze their recipes to find safer alternatives for hazardous substances. Unfortunately, available substance information is dispersed, heterogeneous and stored in databases of many private and public entities. In addition, the number of existing chemical substances already surpassed 85,000 with over 200 attributes describing substance characteristics, which makes it impossible for experts to collect and manually review this data. We tackle these issues by introducing a novel machine learning approach for alternative assessment. After developing a central database, we design an approach that performs nearest neighbor search in latent space obtained by deep autoencoders. Furthermore, we implement a post-hoc explanation technique, t-SNE, to visualize deep embeddings that enables to justify model outcomes. The application in a real-world project with a manufacturer shows that this approach can help process experts to identify possible replacement candidates more quickly and fosters comprehensibility through visualization.

Keywords: worker safety, sustainability, alternatives assessment, deep learning, machine learning

1 Introduction

In the member countries of the European Union, many laws have been installed that aim at reducing the impact that industrial processes may have on the environment and the health of workers and consumers. A prominent example is the RoHS directive (*Restriction of Hazardous Substances*) which, among others, prohibited manufacturer from the use of leaded solder and the sale of any products that contain these regulated substances [1]. Apart from the prohibition to use certain substances, which are determined by the government, legislators within the EU impose companies to take individual action towards reducing the environmental or health-related impact of industrial processes.

As part of these regulations, companies must regularly evaluate the possibilities to substitute hazardous substances used in production with less harmful alternatives. Alternatives assessment (AA) is a process for identifying, comparing, and selecting

14th International Conference on Wirtschaftsinformatik,
February 24-27, 2019, Siegen, Germany

safer alternatives to chemicals of concern (including those in materials, processes, or technologies) on the basis of their hazards, performance, and economic viability [2].

The issue of AA is currently engaged by researchers, government agencies, as well as by NGOs [3, 4]. Government agencies like the European Chemical Agency (ECHA) primarily focus on the management aspect of chemicals by providing data for labeling and classification of substances. Researchers and NGOs, on the other hand, are working on process aspects when conducting AA, resulting in frameworks that support manufacturers in their AA activities [3]. Currently, there are several frameworks for alternatives assessment [3]. At the beginning of an AA process, most researchers suggest the identification of chemicals of concern that will be subject to AA [5–8]. The substances' physio-chemical properties, the human health hazards, and ecotoxicity are subject to evaluation in most of the ten available frameworks identified by [3]. In all frameworks, an economic assessment is part of the AA process after the ecotoxic and health hazard assessment is done [3]. Identifying alternative substances is considered critical in all frameworks, but mostly remains a manual task and is described only on a very vague level. The authors make suggestion on what factors to consider [8], to use market view or literature research for identifying alternative substances [5], or pose questions that should be addressed when trying to find alternative substances [9].

The limitation of available research is threefold. First, most of the existing frameworks are only guidelines for decision making or suited to a certain environment, rather than being a prescriptive protocol to follow. Currently, there is a lack of techniques and information systems (IS) that support companies in finding meaningful alternative substances by addressing some of the issues that come with AA. The second limitation is the complexity of data or the lack of data [3]. The ECHA provides a public database with over 85,000 registered substances on their website [10]. Each data set comes with a classification of the physio-chemical characteristics for each substance using more than 200 different attributes. In addition to the ECHA database, there exist more databases from individual manufacturers (e.g. BASF), trade associations (e.g. GISBAU), governmental agencies, or national social insurance carriers. Since these entities focus on other aspects than ECHA, their data is characterized by other attributes than the ECHA data. This makes it almost impossible for process experts to use this kind of data for AA due to the sheer number of substances and variety of the data, which brings us to the third limitation of current research. Apart from information systems that provide data for individual substances through a web site, there is currently no approach available that helps process experts in extracting and/or condensing meaningful information for their AA activities from existing data.

To the best of our knowledge, there is currently no approach available that specifically supports the identification of finding alternative substances using sophisticated methods from the area of machine learning. We aim at filling this void with the research provided in this paper. We tackle the latter two limitations by designing a software artifact that contains (i) a substance database, which captures the characteristics of the substances that we develop by extracting the data from different sources, (ii) a machine learning (ML) model that examines the substance data to uncover similarities among substances and identifies substitution candidates based on their physio-chemical and hazard-related attributes as well as their functional role [11],

(iii) an explanation interface for ratification of the generated list of substitution candidates. The ML component is based on a nearest neighbor search algorithm combined with the deep stacked autoencoders. For explainability purposes we use the recognized visualization technique, t-SNE, which is assumed to project the obtained autoencoder representations in a two-dimensional space. The resulting IS can potentially help process engineers from different domains to automatically check their bills of materials for the availability of alternative substances that fulfill the same purpose of the original substance, but which are less harmful for the environment, for the workforce, and end users. We demonstrate the applicability of the proposed IS artifact for the production process of a leading German manufacturer of ceramics.

The remainder of the paper is organized as follows: we provide the applied research method in-depth in Section 2. Section 3 introduces the architecture overview and the individual components of the proposed decision support system. More specifically, we explain the database that was created, the ML models that we applied on that data and the explanatory component of the ML technique. Insights on the applicability of our approach are provided in section 4 with a use-case and evaluation in a real company, before we conclude the paper with a summary and a brief overview of future work in section 5.

2 Research Method and Design

The research method applied in this paper shall be characterized as design science research (DSR) [12, 13]. DSR is an important paradigm in IS research and professional practice and serves as a guideline for the process of constructing socio-technical artifacts in the IS domain [14]. The goal of DSR is to solve existing problems with new or improved IS-based solutions [15]. We follow the process of Peffers et al. [13] for creating design science artifacts. The research environment is an interdisciplinary consortium research project that allows us to evaluate our artifacts in a real-world manufacturing situation and collect continuous feedback from project partners [16].

Table 1. Process of creating Design Science Artifacts (adapted from [13])

Step 1: Problem Identification & Motivation	**Step 2: Objectives of a Solution**	**Step 3: Design & Development**	**Step 4: Demonstration**	**Step 5: Evaluation Step 6: Communication**
– Lack of annotated data – Infeasibility of the manual search process – The opaqueness of the black box machine learning techniques	– Collecting data necessary for alternatives assessment – Searching alternative substances with ML techniques – Evaluating the alternatives in terms of ecological and health-related criteria	– Data model and database – ML techniques for reducing complexity of data set – Explanation Interface for ML techniques – Information system for alternative assessment	– Application of the artifacts in the problem context for 4 substances of concern	– Evaluation of the artifacts in a real-world scenario – Presentation of results to academics and to practitioners

At the beginning, the *problem* must be identified and the need for a new or better solution to the problem should be motivated. As mentioned before, the problem in our

case is the current fragmentation of substance data in different databases of individual manufacturers and the lack of methods to process this data [3]. In most current applications such as *P2OASys*[1] or *COSHH*[2], the domain experts must seek for alternative substances manually by analyzing the multidimensional application purposes and the hazard profile of each individual substance, which results in high cognitive load and suboptimal results. Automating the search process with ML techniques is a good alternative, which in turn also requires a careful design and the suitable choice of the approaches. Finally, to embed the ML systems to the decision-making processes it is important to establish the trust to their outcomes, which requires making them explainable by considering the context of application and the user requirements.

In step 2 (*objectives*), the goal of the research must be emphasized. The objective of a solution serves as a baseline of knowledge on which to evaluate the novelty of the created artifacts [14]. The objectives of our approach are to tackle the lack of a sophisticated database with annotated substance data. We do this by extracting necessary data from various sources and creating a single database with harmonized data. We then apply ML techniques to provide means for the expert to evaluate and select alternative substances.

Step 3 (*design & development*) is the phase where the artifacts for the intended solution are created. This can be new constructs, models, methods, or instantiations [12]. We create the design of the database and ML models and implement them as a prototype.

The application of the new artifacts in a suitable context is subject to step 4 (*demonstration*). We chose the context of manufacturing in the ceramics industry to apply our prototype for AA and demonstrate the benefits in this domain.

Step 5 (*evaluation*) is where the final evaluation of the artifacts happens. Based on metrics or observations, the results of the artifacts are evaluated regarding their performance in the field. Since expert knowledge in this step is necessary, we included the ceramics manufacturer to evaluate the resulting list of substitution candidates for five substances that the manufacturer is currently using.

The process of DSR concludes with step 6, where the *communication of the results* to research communities and practitioners takes place.

3 Description of the Proposed Decision Support System

3.1 Architecture overview

With our adapted process (see fig. 1), we follow the framework of the National Research Council [17] for AA. The NRC recommends adopting a multi-perspective approach for AA that begins with the identification of substitution candidates. This is probably the most crucial part of AA due to the information overload for experts. We

[1] see https://p2oasys.turi.org/GetStarted/p2oasys.php

[2] see http://www.hse.gov.uk/coshh/essentials/coshh-tool.htm

support this process with our ML technique, which identifies similar substances, according to their attributes. The physio-chemical attributes must then be assessed by a domain expert to find out, which of the candidates are appropriate for the desired application purpose or product. A comparative exposure assessment for human health hazards as well as ecotoxicity should then be conducted with the remaining substitution candidates. This step is supposed to identify the substance with the smallest negative impact on environment or workforce. We support this step with a web application by providing visualization of the related health or environmental risk for the substitute candidates. Depending on the application environment, the process expert can weigh the attributes differently. For example, in a scenario where water is drained off to the environment after usage, a substance with less risk for causing aquatic toxicity would be more appropriate whereas a substance whose dust can be harmful to the workforce would be less problematic in a fully automated process. Finally, economic considerations like market prices or availability of substances are considered. These three steps are recommended by the majority of established AA frameworks [17].

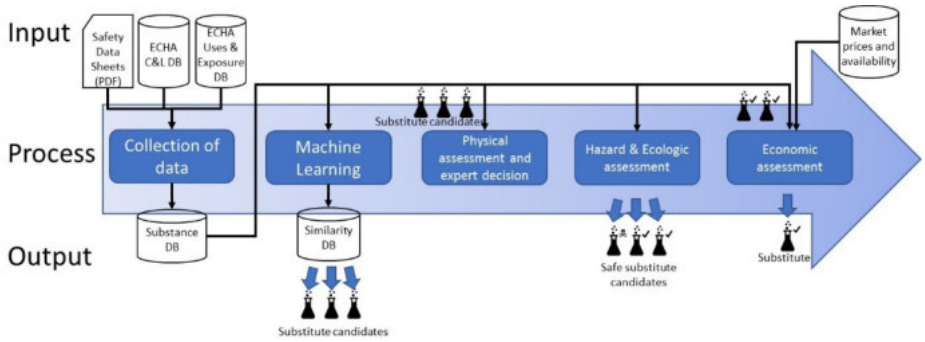

Figure 1. Architecture of the information system

Since the related research project is ongoing and we focus on environmental and health issues in this paper, we present the results up to the hazard & ecologic assessment and focus on the ML component in the following sections.

As mentioned in the introduction, there is currently a need for IS that support manufacturers in comparing available substances, e.g. for planning and maintaining product design. This IS must be capable of analyzing large sets of chemical substance data using ML models. Since many manufacturers have no data science experts available, the results of the ML techniques must be provided in a way that can easily be understood by domain experts.

The solution that we propose is based on a web application that we implemented using Java Enterprise for the user interface and business logic, and R statistical computing programming language for the ML components. In the end user perspective, the user can start a new AA process by providing the CAS# of the substance that is subject to the assessment along with meta data for the AA project. Thereafter, the user can choose how many possible substitutes should be presented and if potentially dangerous substances from the list of substances of very high concern (SVHC) which

is maintained by ECHA should be included in the suggestions. The user will then be provided with the interactive scatter chart from Figure 3 with the initial substance highlighted. The samples near the examined instance are provided as colored dots depending on the number of suggestions that was selected. Once the users click these points, these substances will immediately be added to a table below the chart that shows the physio-chemical attributes of every selected substance. After a manual review of the relevant attributes, the user can select the remaining candidates for the next process step. In the next step, the remaining suggestions will be compared among each other regarding their hazard for human and environment. This is done using the GHS column model [18], where the hazard of each substance is benchmarked regarding six different dimensions (see fig. 4).

The administrator of the web app can access the machine learning perspective to start a new learning phase. During the learning, the substance data is retrieved from the database, the autoencoders and t-SNE models are applied to identify and to visualize the nearest neighbors for each substance. The results in the form of an x- and y-value for the position of each substance in a 2d space are stored in the database. The learning must be recomputed as soon as new substances considered are added to the database.

3.2 Database Component: Developing a Central Database

The design of a comprehensive database that can be used for AA in manufacturing companies requires collecting and processing available data. The objective of the ECHA is to collect data about hazardous substances that are imported to or produced in the European Union. ECHA is constantly assessing the chemical-related risks of many substances through manufacturer provided dossiers, own tests, and chemical analyses. Thus, ECHA has available a large volume of data, that is publicly available through their web site.

Most important for identifying initial substitute candidates is data regarding the physical attributes of substances. This kind of data is provided by ECHA through their *registered substances list*, where information about the use and exposure (U&E) of substances is available. The U&E data is classified by the following use descriptors.

Table 2. Attributes of the uses and exposure data set

Attribute type	**Sector of use** (SU, 24 attributes)	**Process category** (PROC, 31 attributes)	**Product category** (PC, 45 attributes)	**Article Category** (AC, 81 attributes)	**Environmental Release Category** (ERC, 26 attributes)
Description	The sector in which the substance is used (e.g. industrial, consumer)	The application techniques or process types	Types of chemical products in which a substance is used.	Type of article in which the substance has been processed.	Broad conditions of use from an environmental perspective.

A second type of data available from ECHA is the *classification and labelling inventory* (C&L). The C&L provides electronic public access to possible hazards for workforce and environment that can be caused by a substance. This data is important for the hazard and ecologic assessment step. The basis for the characterization is the

Globally Harmonized System of Classification and Labelling of Chemicals (GHS) with its 100 attributes in total. These binary-coded attributes describe, whether a substance is causing a certain physical, health, or environmental hazard or not.

As a third source for acquiring substance data we implemented a software tool for extracting relevant information from existing safety data sheets (SDS) in the PDF format. These sheets provide a standardized set of data regarding substance meta data (Manufacturer, retail name), composition/information on ingredients, or relevant identified uses of the substance or mixture. The matching between the data extracted from SDS and the C&L inventory was made through the chemical abstract number (CAS#), a unique identifier for every substance which is available at every SDS and in the C&L and U&E inventory.

Through these three data sources, we could collect 13,133 data instances regarding U&E of substances and 85,845 data instances of the C&L inventory.

3.3 Modelling Component: Nearest Neighbor Search and Deep Autoencoders

3.3.1 Finding Alternative Substances with Nearest Neighbor Search

Once the central database is developed, we aim to propose an intelligence component that presents the production experts the most suitable alternatives to the substances that are used in the current production processes. For this purpose, the most similar matches to high dimensional query vector that captures different aspects of the usage- and hazard profile can be extracted by using the nearest neighbor search algorithm. However, searching the similar items in large datasets is a challenging issue. Hence, the suitability of the retrieved neighbor instances via nearest neighbor search algorithm is affected negatively by the inappropriate adoption of distance/similarity measure and existence of various class irrelevant features [19]. Furthermore, the process of measuring the similarities between the query case and historical items with high feature dimensions from very large validation sets is very time consuming and suffers from the curse of dimensionality which makes infeasible to identify the exact matches in reasonable computational costs [20].

To alleviate these obstacles, various techniques have been proposed and extensively investigated [20]. Neighborhood identification in the latent space obtained by applying dimensionality reduction techniques is considered as the most promising approaches among others [21]. Earlier studies in this field investigated diverse linear approaches such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA) among others to map the original high dimensional feature space to the low dimensional latent space linearly [19]. However, such a linear transformation shows an inability to model the higher-order correlation in the original data space and is deficient in capturing the intrinsic class-specific data manifold [19, 22]. On these grounds, we can argue that searching for similar instances in the latent space obtained by linear transformation may result in the retrieval of the inappropriate explanation items.

Different non-linear transformation and dimensionality reduction approaches have been proposed to surmount the problems arising from these shortcomings. These techniques are assumed to place the items having the same characteristics near to each other in the non-linearly transformed low-dimension feature space [19]. In our study, we employ the stacked autoencoder based deep neural networks, which extract features that constructs useful higher-level representations. The empirical experiment results on various datasets by [22, 23] provide confirmatory evidence that using the latent representations obtained by autoencoders shows superior performance compared to linear PCA, LDA and other non-linear approaches such as linear embeddings in retrieving similar items.

3.3.2 Autoencoder Neural Codes for Nearest Neighbor Search

Autoencoders is the type of unsupervised feed-forward neural networks with three layers namely, an input layer, a hidden layer and an output layer in which the training purpose is defined as reproducing the input data at the output layer. The component of autoencoder networks which computes the hidden layer activations from the input data are referred as encoders. The encoder employs a non-linear mapping function to obtain the hidden layer representation. Following this, the decoder component maps the latent representation of the input data obtained and presented in the hidden layer through the encoding process to a reconstructed vector in high-dimensional input space by using the chosen non-linear mapping function with reverse mapping parameter set. The parameter sets of both encoding and decoding layer are optimized simultaneously with the aim to minimize the average reconstruction error.

Deep neural networks can be built by stacking multiple layers of autoencoders which have already been trained locally as described in the previous section [24]. The training process of the stacked autoencoders with k layers is initialized by training first autoencoder in which the original input data are used both in input and output layers. The learned hidden layer activations are then wired to input and reconstruction layers of the second autoencoder with the purpose to obtain the corresponding hidden layer features and parameters. This greedy layer-wise learning is performed until all features including the ones from last k^{th} hidden layer are extracted and all relevant parameters, weights, and biases are initialized. An example training process of deep neural networks with two hidden layers is depicted in Figure 2.

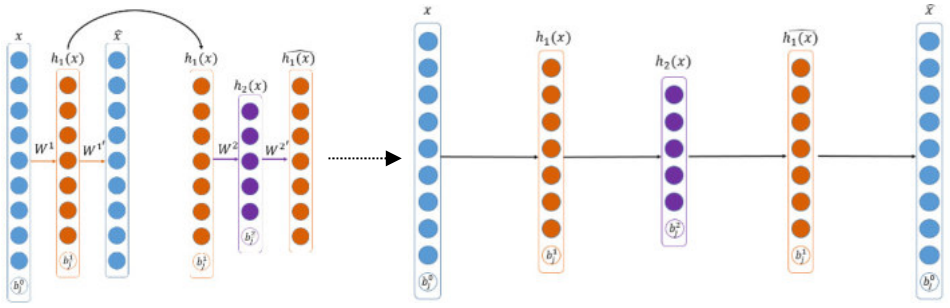

Figure 2. Layer-wise training of stacked autoencoders

As described above the main idea of (stacked) autoencoders is to learn the useful latent representations by teaching the network to copy the input to output. For this purpose, it is very important to add either architectural constraints or employ the regularization techniques during the learning process. In the last decade various autoencoder architectures such as stacked denoising autoencoders, variational autoencoders, contractive autoencoders, regularized autoencoders, undercomplete autoencoders and others have been proposed which are assumed to increase the quality of the extracted latent features. Since we aim in our study also to conduct the nearest neighbor search process based on hidden layer activations faster and more efficiently, it is reasonable to adopt an autoencoder architecture with the decreasing width (layer sizes) of hidden layers, namely undercomplete autoencoders. The main idea behind this type of the autoencoders is to constrain the number of nodes in the hidden layer and to force the model to learn the most important features form the inputs while minimizing the reconstruction error.

After mapping all substances in the dataset onto useful latent space with autoencoders we can find the similar instances to the given query instance by searching its neighbors in the latent space. For this purpose, we first extract the bottleneck features of the query instance by feeding its original input values to the learned deep autoencoder, by computing the hidden layer activations and by calculating the distances between them and the bottleneck features of all other instances. The list of substances with the smallest distance to the query instance is presented to the experts, which are later examined whether they are sustainable alternatives in terms of productions and economic factors.

3.4 Explanation Component: Visualization with t-SNE

In order to exploit the full potential of ML techniques, their outcomes have to be embedded to the production and business processes that creates value by extending the corporate ability to gain new insights [25]. However, a series of recent studies have indicated that the lack of trust in the machine learning model with opaque reasoning mechanisms is considered as one of the main obstacles in operationalizing the data driven analytics that in turn contributes to the extending gap in the scientific developments and their practical applications [26, 27]. Making the advanced black-box models such as deep learning techniques implemented in the current study or their outcomes explainable is considered a potential solutions that has recently received substantial interest [28]. The practicability and reliability of the explanations depend significantly on the context of explanation situation and the target audience that must be considered when choosing and designing the relevant techniques. In our use case, the machine learning explainability is required for the end users, namely the production domain experts, who have little interest in understanding the reasoning procedure of the implemented nearest neighbor search in latent space obtained by deep stacked autoencoders [29]. With this in mind, the post-hoc explanation family techniques are

suitable for this purpose since they are assumed to increase the end users' confidence to the ML models by allowing to justify the outcomes delivered by the system [30].

Since the proposed ML model in this study provides the end users the list of similar alternatives to the query case, it already incorporates an intrinsic example-based explanation ability. The domain experts can ratify the suitability of the generated alternatives list by using their domain knowledge and evaluate their validity without a need for examining how and why it was generated by the system. However, such explanations facilitate the users to understand the model locally by providing the list of alternatives for one substance at a time. To enable the users to analyze and understand the entire model outcomes in the whole instance space we adopt another post-hoc explanation technique, visualization with t-Stochastic Neighborhood Embeddings (t-SNE). t-SNE is considered as one the most powerful dimensionality reduction and visualization techniques, that can effectively visualize the high dimensional data by assigning each data point a location in a two or three dimensional map [31]. The main superiority of the t-SNE is its ability to visualize the similarity data by retaining the local structure of the data while also providing the relevant information about the global structure. It is important to note that we visualize the embeddings obtained by the compression of the original data via deep stacked autoencoders with an extension of the original t-SNE technique, namely its implementation with Barnes-Hut approximation algorithm. Implementing the later approach leads to substantial computational efficiency by enabling to embed the high dimensional data with the runtime $O(N \log N)$ instead of $O(N^2)$.

4 Experimental Settings and Results

4.1 Use Case: Substitution of Hazardous Substances in Ceramics Industry

The research covered by this paper was conducted in the ceramic industry. The company that served as partner is a German manufacturer of sanitary ceramics with manufacturing facilities in different European countries. The company uses a variety of chemical substances and mixtures during the three steps of manufacturing. At the very beginning, substances are used, for example, as adhesive or cleansing agent for the construction of the molding tools (i.e. the negative). Then the compound for molding is mixed in a large pug mill from several substances. After the mix is ready, the molding tool is used for high-pressure injection molding of the mix which forms the workpiece. In a series of further manual steps, grinding and glazing of the workpiece takes place. In each of these steps, workers can be exposed to hazardous substances through skin contact or inhale of dust. Most substances have a wide range of short and long-term effects, ranging from skin irritation to carcinogenicity. To reduce the risk for workforce and environment, the firm is looking for new techniques of AA that are robust and can cope with the challenges that massive amount of data present today.

4.2 Tools and Model Parameters

All models for machine learning and explanation components of the proposed solution were developed by using the R statistical computing programming language. The stacked autoencoders were implemented on top of the h2o package [32]. The input layer consists of 208 nodes and the bottleneck code layer size was defined as 20. To avoid the overfitting, we adopted the early stopping technique in which the Mean Squared Error was used as stopping criterion. All layers of the implemented stacked autoencoders used the Tanh activation function. For nearest neighbor search we used the brute force nearest neighbor search in the latent space obtained by stacked autoencoders by computing the Euclidean distance between the query case and all other substances in the dataset. For this purpose we used the FNN package [33]. Performing the nearest neighbor search in the latent space has resulted in an acceleration by factor 11 compared to the search in original high dimensional space.

Table 3. t-SNE and stacked autoencoder parameters

t-SNE		Autoencoder	
Parameter	Value	Parameter	Value
Dimension of Embedded Space	2	Layers	208-50-20-50-208
Learning Rate	200	Activation	Tanh
Perplexity	30	Epochs	1000
Iterations	1000	Learning Rate	0.005
Initial Momentum	0.5	Rate Annealing	0.000001
Momentum Switch Iteration	250	Initial Momentum	0.5
Final Momentum	0.8	Final Momentum	0.99
Theta (Speed/Accuracy Trade-off Parameter)	0.5	Rate Decay	1
Early Exaggeration	12	Initial Weight Distribution	UniformAdaptive

Following this, we visualized all data represented in the bottleneck neural codes of the implemented autoencoder by using the t-SNE with Barnes-Hut approximation algorithm and follow the training instructions by its developers to obtain reliable results [34]. Particularly, a gradient descent optimization was performed for 1000 iterations. Furthermore, an early exaggeration trick was implemented when minimizing the KL-divergence. An additional momentum term of 0.5 was used for the first 250 iterations whereas it was increased to 0.8 for the rest. To perform t-SNE with Barnes-Hut approximation, we used the Rtsne package [35]. A comprehensive overview of parameters for deep stacked autoencoders and t-SNE are depicted in Table 3.

4.3 Demonstration and Evaluation

To demonstrate the contribution of our approach to AA, we examined four different substances that are currently being used in the production process of our industry partner, that are considered harmful for workers or the environment. These substances were selected by the ceramic manufacturer due to their individual risk profile: *Barium Carbonate, Instapak A, Sodium Borate,* and *Hydrogen Peroxide 30%*. We identified ten substitutes for every substance based on the proposed nearest neighbor search

method (see fig. 3). The overall impression after an initial review of the ten suggestions by the expert was that for three of the four initial substances, at least one reasonable substitution candidate was provided through the machine learning process. Only in the case of *Instapak A*, substances cannot be easily exchanged since *Instapak A* is used as one component in a multi-component ready-to-use packaging foam provided by a third party. In other cases, a further investigation is necessary to back the positive first impression through professional literature and studies. An excerpt for the substance *Barium Carbonate* and the first five most important substitute candidates based on nearest neighbor search is listed in Table 4. We will explain the results of the evaluation in the case of *Barium Carbonate* in more detail. *Barium Carbonate* acts as a network converter in the glazing agent to increase the viscosity of the glaze mix in the fire kiln. It is also lowering the melting temperature of the glaze. At first sight, it is obvious that the first substitute candidate is related to the original substance based on its name. The base material of *Barium Carbonate* is *Barium*. The ML algorithm identified *Barium* as a substitute candidate for *Barium Carbonate* because the physio-chemical attributes of the two are almost identical. The second substitute, *Boron Nitride*, turned out to be impractical from the standpoint of technical product design. *Lead*, on the other hand, is a member of the SVHC list provided by ECHA and is even more dangerous than *Barium Carbonate*. Therefore, it was not considered in the further process. With *Carbonic Acid* we discovered a substance that fits the spectrum of substitute candidates from a technical, and safety standpoint. Thus, we proceeded with that substance to the next AA process step. What the evaluation of the first process step revealed is that the final choice of substitute candidates cannot be made by a data scientist or an algorithm alone. Instead, a domain expert from the manufacturing company must conduct a preliminary assessment to choose, which substitute candidates should be regarded in the further evaluation process. We also received the feedback that a function to exclude substances from the SVHC list from the suggestions would reduce the effort necessary for the manual review of the results.

Table 4. First five replacement substances for barium carbonate

Orig. substance	1st substitute	2nd substitute	3rd substitute	4th substitute	5th substitute
Barium carbonate CAS# 513-77-9	Barium CAS# 7440-39-3	Boron nitride CAS# 10043-11-5	Dicopper oxide CAS# 1317-39-1	Lead CAS# 7439-92-1	Carbonic acid CAS# 463-79-6

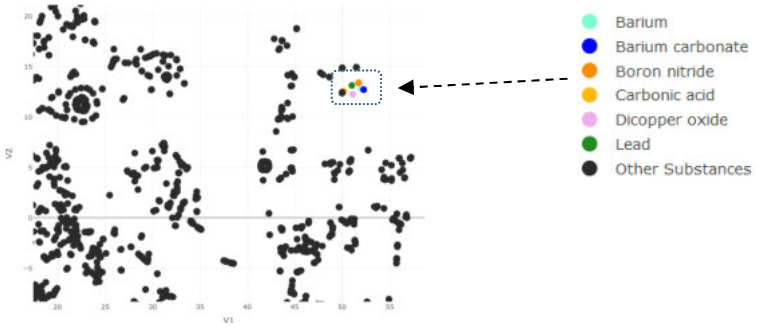

Figure 3. Visualization of substances with t-SNE (zoomed region)

The second step of our AA process was the comparison of the hazard and ecological factors that the new substances raise. Depending on the application context, a substitution should only take place, if the relevant risks for human and environment decrease due to the substitution. To compare the risks of both substances, we applied the GHS column model [18], which helps us to compare the two substances regarding the six criteria listed as columns in Figure 4.

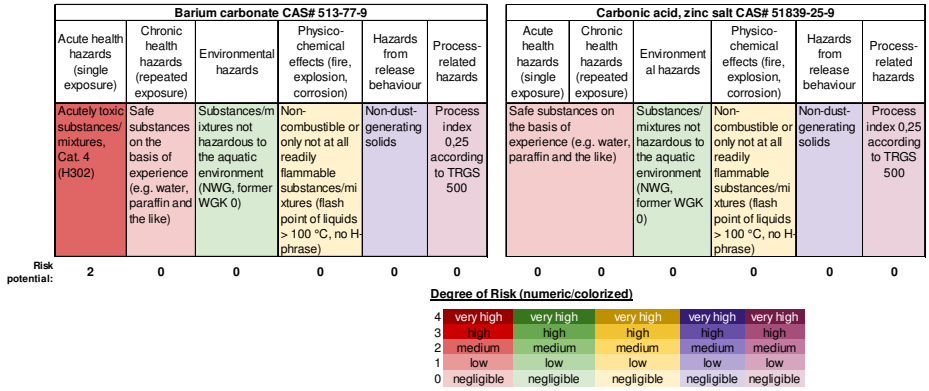

Figure 4. Hazard related comparison between substances

We compared *Barium Carbonate* and *Carbonic Acid*, which was ranked 5th on the substitute candidate list. As can be seen in Figure 4, *Barium Carbonate* and *Carbonic Acid* have an almost identical hazard profile. Acute health hazards that unfold at every exposure for *Carbonic Acid* pose only a low risk compared to a high risk of *Barium Carbonate*. That makes *Carbonic Acid* the substance of choice for manual tasks with substance exposure to the workforce.

5 Discussion of the Results and Conclusion

As ML is gaining momentum in many fields, new opportunities for solving information intensive tasks using ML techniques emerge. With alternatives assessment, we applied ML on an information intensive task. As emphasized before, the overload on information when it comes to assessing alternative substances along with the attitude to "never change a winning team" is what is causing many lost opportunities to transform production processes or recipes to more environmentally or workforce friendly variants.

The results of our research show that AA is a prime example for applying ML to ease existing problems, that are often not properly addressed due to lack of experts or due to resistance to change. Surveys reveal that *lack of information on alternatives* along with a *lack of relevant expertise and resources in companies* are two major inhibitors when it comes to AA [36]. We addressed the first issue by providing an IS and a database along with the identification of relevant data that fosters alternatives assessment. To approach the second problem, we proposed an ML technique that

supports domain experts in finding reasonable substitution candidates that can be further assessed in terms of their economic aspects. The feedback that we received from our project partner so far was promising and the results are further evaluated by the product engineering department.

While the research project that is subject to this paper is still ongoing, future efforts shall focus on further aspects of AA. For example, on the integration of enterprise IS to improve the process of AA. Currently, each substance must be assessed manually in our application. Providing data of recipes directly from ERP systems would allow us to process large sets of data and find similar but less toxic substitute candidates on a broader scale. In addition, the economic aspect of a substitution will be focused on and added to our application.

6 Acknowledgment

The research described in this paper was supported by a grant from the German Ministry of Education and Research (BMBF), project name SiFa 4.0 (01IS17009B). Parts of the data was kindly provided by the European Chemicals Agency (ECHA).

References

1. Cusack P, Perrett T (2006) The EU RoHS Directive and its implications for the plastics industry. Plast Addit Compd 8:46–49
2. BizNGO (2013) The Commons Principles for Alternatives Assessment: Addressing Chemicals of Concern to Human Health or the Environment, https://www.turi.org/TURI_Publications/TURI_Chemical_Fact_Sheets/Commons_Principles_for_Alternatives_Assessment (Accessed: 19.07.2018)
3. Jacobs MM, Malloy TF, Tickner JA, Edwards S (2016) Alternatives assessment frameworks: Research needs for the informed substitution of hazardous chemicals. Environ Health Perspect 124:265–280
4. Geiser K, Tickner J, Edwards S, Rossi M (2015) The Architecture of Chemical Alternatives Assessment. Risk Anal 35:2152–2161. https://doi.org/10.1111/risa.12507
5. Rossi M, Tickner J, Geiser K (2006) Alternatives Assessment Framework of the Lowell Center for Sustainable Production. http://www.sustainableproduction.org/downloads/FinalAltsAssess06_000.pdf (Accessed: 21.08.2018)
6. Rossi M, Peele C, Thorpe B (2012) BizNGO Chemical Alternatives Assessment Protocol, https://www.biznGO.org/alternatives-assessment/chemical-alternatives-assessment-protocol (Accessed: 15.08.2018)
7. U.S. EPA (2011) Design for the Environment Program Alternatives Assessment Criteria for Hazard Evaluation, https://www.epa.gov/sites/production/files/2014-01/documents/aa_criteria_v2.pdf (Accessed: 25.08.2018)
8. Reihlen A (2011) Guide on Sustainable chemicals: decision tool for substance manufacturers, formulators and end users of chemicals, https://www.subsport.eu/wp-content/uploads/2011/10/UBA_guide_sustainalbe_chemicals.pdf (Accessed: 28.07.2018)
9. IC2 (2013) Alternatives Assessment Guide, http://theic2.org/article/download-pdf/file_name/IC2_AA_Guide_Version_1.1.pdf (Accessed: 30.07.2018)
10. Erickson B (2017) How many chemicals are in use today? Chem Eng News 95:23–24

11. Isaacs KK, Goldsmith MR, Egeghy P, et al (2016) Characterization and prediction of chem. functions and weight fractions in consumer products. Toxicol Reports 3:723–732
12. Hevner AR, March ST, Park J, Ram S (2004) Design Science - Hevner. Des Sci Inf Syst Res Author MIS Q 28:75–105
13. Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S (2007) A Design Science Research Methodology for Information Systems Research. J Manag Inf Syst 24:45–77
14. Gregor S, Hevner AR (2013) Positioning and presenting design science research for maximum impact. MIS Q 37:337–355
15. Winter R (2008) Design science research in Europe. Eur J Inf Syst 17:470–475
16. Österle H, Otto B (2010) Consortium Research. Bus Inf Syst Eng 2:283–293
17. National Research Council (2014) A Framework to Guide Selection of Chemical Alternatives. Nat. Acad. Press, http://www.nap.edu/catalog/18872 (Accessed: 10.08.2018)
18. Smola T (2014) The GHS Column Model as an aid to selecting substitute substances, https://www.dguv.de/medien/ifa/en/pra/ghs_spaltenmodell/spaltenmodell_2017_en.pdf (Accessed: 12.07.2018)
19. Min R, Stanley DA, Yuan Z, et al (2009) A deep non-linear feature mapping for large-margin knn classification. Ninth IEEE Int. Conference on Data Mining, 2009. pp 357–366
20. Muja M, Lowe DG (2014) Scalable nearest neighbor algorithms for high dimensional data. IEEE Trans Pattern Anal Mach Intell 36:2227–2240
21. Keogh E, Chakrabarti K, Pazzani M, Mehrotra S (2001) Locally adaptive dimensionality reduction for indexing large time series databases. ACM Sigmod Rec 30:151–162
22. Salakhutdinov R, Hinton G (2007) Learning a nonlinear embedding by preserving class neighbourhood structure. In: Artificial Intelligence and Statistics. pp 412–419
23. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science (80-) 313:504–507
24. Larochelle H, Bengio Y, Louradour J, Lamblin P (2009) Exploring strategies for training deep neural networks. J Mach Learn Res 10:1–40
25. Davenport TH, Harris JG (2007) Competing on analytics : the new science of winning, 1st ed. Harvard Business School Press
26. Sanders NR, Manrodt KB (2003) The efficacy of using judgmental versus quantitative forecasting methods in practice. Omega 31:511–522
27. Halper F (2016) Operationalizing and Embedding Analytics for Action, https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/tdwi-operationalizing-embedding-analytics-for-action-108112.pdf (Accessed: 20.07.2018)
28. Gunning D (2017) Explainable Artificial Intelligence (XAI). DARPA
29. Gregor S, Benbasat I (1999) Explanations from intelligent systems: theoretical foundations and implications for practice. MIS Q 23:497–530
30. Lipton ZC (2016) The Mythos of Model Interpretability. ArXiv e-prints, June 2016.
31. Van der Maarten L., Hinton G (2008) Visualizing Data using t-SNE . Journ of Mach Learn Res 9 (2008) 2579-2605
32. Candel A, Parmar V, LeDell E, Arora A (2016) Deep learning with h2o. H2O Inc
33. Beygelzimer A, Kakadet S, Langford J, et al (2013) FNN: fast nearest neighbor search algorithms and applications. R Packag version 1
34. Wattenberg M, Viégas F, Johnson I (2016) How to use t-sne effectively. Distill 1:e2
35. Krijthe JH (2015) Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. R Packag version 013, URL https//github com/jkrijthe/Rtsne
36. Tickner J, Jacobs M (2016) Improving the identification, evaluation, adoption and development of safer alternatives: Needs and opportunities to enhance substitution efforts within the context of REACH. Environ Health Perspect. 2016 Mar;124(3):265-80